

**Lieff
Cabraser
Heimann &
Bernstein**
Attorneys at Law

Susman Godfrey I.L.P.
a registered limited liability partnership



November 19, 2024

VIA ECF

Hon. Ona T. Wang
Daniel Patrick Moynihan
United States Courthouse
500 Pearl St.
New York, NY 10007

RE: *Authors Guild v. OpenAI Inc.*, 23-cv-8292 (S.D.N.Y.) and *Alter v. OpenAI Inc.*,
23-cv-10211 (S.D.N.Y.): Microsoft's LLMs and Licensing

Dear Judge Wang:

Pursuant to Rule II(b) of Your Honor's Individual Practices, Plaintiffs seek a conference regarding Microsoft's refusal to produce two categories of information on the ground of relevance: **(1)** documents relating to Microsoft's efforts to license LLM training data where a license did not result (RFP Nos. 46 and 47) (Ex. 1) and **(2)** documents showing the datasets that Microsoft used to train its LLMs (RFP No. 41) (*id.*).¹ Both categories are relevant and discoverable. Negotiations and draft licensing deals for LLM training data—consummated or not—are relevant to the existence of a potential or actual licensing market for training data under the fourth fair use factor. Information about how Microsoft trained its LLMs is relevant to both Plaintiffs' direct *and* contributory infringement claims against Microsoft.

Background. RFP Nos. 46-47 seek Microsoft's negotiations to license copyrighted works (Ex. 1). To date, both party and non-party discovery confirms that Microsoft sought to license training data from [REDACTED] (Exs. 2, 3). While Microsoft agreed to produce some licensing material, it has – unlike OpenAI – categorically refused to produce documents relating to (1) unexecuted licenses; or (2) licensing training data for its own LLMs (*see* Ex. 4).

RFP No. 41 seeks documents sufficient to identify the datasets that Microsoft used to train LLMs. In its response to this RFP and several others, Microsoft objected to producing documents related to its own LLMs (Ex. 5). Microsoft reiterated these objections during the parties' conferrals. On November 15, 2024, Microsoft confirmed that it would not change its position as to the requests (*see* Ex. 4) ("The allegations in the amended complaint . . . do not extend to LLMs other than OpenAI's LLMs").

Plaintiffs are suing on behalf of a class of people who own "Class Works": works that "ha[ve] been, or [are] being, used by Defendants to 'train' one or more of [their] large language models." Compl. ¶¶ 394, 397. On behalf of that class, Plaintiffs assert direct and contributory infringement claims against Microsoft. Paragraph 415 of the direct infringement claim alleges that

¹ The parties have conferred and were not able to reach agreement.

November 19, 2024

Page 2

“Defendants . . . reproduc[ed] the works owned by Plaintiffs . . . to train their artificial intelligence models.” Among those and other allegations, Plaintiffs also plead:

- “Defendants OpenAI and Microsoft have enjoyed enormous financial gain from their free exploitation of copyrighted material.” ¶ 9.
- “Neither OpenAI nor Microsoft have paid for the books used to train their models. Nor have Defendants sought to obtain—or pay for—a license to copy and exploit the protected expression contained in the copyrighted works used to train their models. Instead, Defendants took these works; they made unlicensed copies of them.” ¶ 10.
- “Defendants have acted on grounds common to Plaintiffs and the Classes by treating all Plaintiffs’ and Class Members’ works equally, in all material respects, in their LLM ‘training.’” ¶ 406.

A. Licensing communications and negotiations (RFP Nos. 46-47).² Microsoft’s licensing discussions are highly relevant to various issues, including fair use, damages, and willfulness. And these discussions and negotiations are no less relevant even where they did not result in a license.

Existence of an actual or potential licensing market. The fourth fair use factor concerns the “effect of the use upon the potential market for or value of the copyrighted work.” 17 U.S.C. § 107(4). In considering whether a market is relevant to fair use, courts consider whether the market is “traditional, reasonable, or likely to be developed.” *Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 929–30 (2d Cir. 1994). A key relevant market here is the licensing market for LLM training data. Whether such an actual or potential licensing market exists is not just evidenced by executed agreements. Microsoft’s efforts to negotiate a license for LLM training data show the existence of demand for the material, and the existence—or at a minimum, the development—of a licensing market.

Damages. Licensing-related communications and negotiations also bear directly on damages. The reasonable licensing fee a copyright owner could have obtained is considered in the calculation of both statutory and actual damages. *See, e.g., On Davis v. The Gap, Inc.*, 246 F.3d 152, 164–68 (2d Cir. 2001) (actual damages); *Reilly v. Commerce*, 2016 WL 6837895, at *8-9 (S.D.N.Y. Oct. 31, 2016) (statutory damages). Prices discussed in licensing negotiations go directly to the value of a license, to lost revenues, and ultimately to damages. *Ringgold v. Black Entm’t TV, Inc.*, 126 F.3d 70, 81 (2d Cir. 1997) (finding a market where plaintiff “was asked to license use of [her work] by the producers of another TV sitcom and declined because of an inadequate price”). In fact,

Ringgold, 126 F.3d at 81.

Willfulness. If Microsoft engaged in licensing negotiations, that is evidence that it knew licensing was required and that its infringement was thus willful. *Broad. Music*, 158 F. Supp. 3d at 197. Again, that is no less true if negotiations fell through: if Microsoft determined that it needed a license, was not able to obtain one, and still proceeded with training LLMs on that data, that is

² These arguments also apply to Microsoft’s objections to ROG 4-5 and warrant compelled responses.

November 19, 2024

Page 3

highly relevant to whether its infringement was willful.

B. Microsoft's own LLMs (RFP No. 41). Microsoft's contention that it does not have to produce discovery about the data it used to train its own LLMs is wrong for two reasons. First, those models are a predicate for Plaintiffs' direct infringement claim against Microsoft. If those models infringed Plaintiffs' copyrights, Microsoft is liable to Plaintiffs. Second, the data Microsoft chose for training its models is relevant to its knowledge of OpenAI's infringing conduct and its own willfulness.

Microsoft's models are part of Plaintiffs' claims. As described, Microsoft's LLMs are within the scope of Plaintiffs' claims. Plaintiffs allege a direct infringement claim against Microsoft, and, in several places, expressly allege that "*Defendants*"—including Microsoft—committed copyright infringement by using authors' works to train "*Defendants*' models." See Compl. ¶¶ 9-10, 406, 415. Indeed, Plaintiffs define their class as those whose works were improperly used to train any of "*Defendants*' large language models." *Id.* ¶¶ 393-397. That Microsoft ignores allegations about its own LLMs—and the class definition—is no basis to limit the scope of discovery. See, e.g., *XChange Telecom Corp. v. Sprint Spectrum L.P.*, 2015 WL 773752, at *3 (N.D.N.Y. Feb. 24, 2015) ("No party possess[es] the unilateral ability to dictate the scope of discovery based on their own view of the parties' respective theories of the case.") (internal quotation marks omitted). Plaintiffs have pleaded that Microsoft's LLMs infringe Plaintiffs' copyrights, and Plaintiffs are entitled to find out what data Microsoft's LLMs are trained on.

Microsoft's own model training is relevant to its knowledge and willfulness. Parties "may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense and proportional to the needs of the case." Fed. R. Civ. Proc. 26(b)(1). Relevancy "is an extremely broad concept," and there is a "relatively low threshold for a party to show that the material sought is relevant to any claim or defense in the litigation." *Delta Air Lines, Inc. v. Lightstone Grp., LLC*, No. 21-MC-374 (RA) (OTW), 2021 WL 2117247, at *2 (S.D.N.Y. May 24, 2021). Material relating to Microsoft's LLMs bears directly on the issues of knowledge and willfulness. Knowledge of OpenAI's infringement is an element of Plaintiffs' contributory infringement claim against Microsoft. See, e.g., *White v. DistroKid*, 2024 WL 3195471, at *10 (S.D.N.Y. June 24, 2024). Willfulness, meanwhile, is relevant to damages: the statutory damages cap for willful infringement is far higher than that for non-willful infringement.

Evidence that Microsoft refrained (or did not refrain) from training its own LLMs on copyrighted material is probative of whether Microsoft knew—or was turning a "blind eye" to the fact, *Arista Records LLC v. Usenet.com, Inc.*, 633 F. Supp. 2d 124, 154 (S.D.N.Y. 2009), that OpenAI's unauthorized use of copyrighted material to train LLMs infringed. That awareness is an element of Plaintiffs' claim for contributory infringement. *DistroKid*, 2024 WL 3195471, at *10 (contributory infringement requires showing "that a party with "knowledge of the infringing activity[] induce[d], cause[d], or materially contribut[ed] to the infringing conduct of another.") (internal quotation omitted). It also goes to Microsoft's willfulness. *Bryant v. Media Rights Prods., Inc.*, 603 F.3d 135, 143 (2d Cir. 2010) (willfulness may involve "knowledge that its conduct represented infringement or recklessly disregarded the possibility"); *Broad. Music*, 158 F. Supp. 3d at 197 (willfulness "is often found where a defendant continued the infringing behavior" knowing that a license was required).

November 19, 2024

Page 4

Sincerely,

LIEFF CABRASER HEIMANN
& BERNSTEINS LLP

SUSMAN GODFREY LLP

COWAN, DEBAETS,
ABRAHAMS & SHEPPARD
LLP

/s/ Rachel Geman

Rachel Geman

/s/ Rohit Nath

Rohit Nath

/s/ Scott J. Sholder

Scott J. Sholder